ISSN 1975-8359(Print) / ISSN 2287-4364(Online)
The Transactions of the Korean Institute of Electrical Engineers Vol. 66, No. 8, pp. 1265~1271, 2017
http://doi.org/10.5370/KIEE.2017.66.8.1265

특징 선택과 융합 방법을 이용한 음성 감정 인식

Speech Emotion Recognition using Feature Selection and Fusion Method

김 원 구*
(Weon-Goo Kim)

Abstract - In this paper, the speech parameter fusion method is studied to improve the performance of the conventional emotion recognition system. For this purpose, the combination of the parameters that show the best performance by combining the cepstrum parameters and the various pitch parameters used in the conventional emotion recognition system are selected. Various pitch parameters were generated using numerical and statistical methods using pitch of speech. Performance evaluation was performed on the emotion recognition system using Gaussian mixture model(GMM) to select the pitch parameters that showed the best performance in combination with cepstrum parameters. As a parameter selection method, sequential feature selection method was used. In the experiment to distinguish the four emotions of normal, joy, sadness and angry, fifteen of the total 56 pitch parameters were selected and showed the best recognition performance when fused with cepstrum and delta cepstrum coefficients. This is a 48.9% reduction in the error of emotion recognition system using only pitch parameters.

Key Words : Emotion recognition, Pitch, Parameter fusion, Feature selection

1. 서 론

정서적 능력은 인간 지능의 중요한 상징이며, 대인 관계에서의 감정은 필수적이다. 인간은 인간의 언어를 이해하고, 사람의 감정을 판단하며, 보다 자연스럽고 조화로운 인간-컴퓨터 상호 작용을 실현하기 위해 노력해 왔다. 현재 음성 인식과 화자 인증 기술은 실용화가 가능할 정도로 안정화되고 있고, 미래기술로 많은 관심과 주목받고 있는 분야는 감정인식 분야이다. 특히, 기계가 인간 감정을 파악하고, 그에 따라 정서적으로 반응하는 시스템 개발이 개발된다면 보다 고차원적인 인간-컴퓨터 인터페이스 제품 개발과 고객 중심의 맞춤형 서비스가 가능하게 될 것이다. 이에 따라 유럽, 미국, 일본, 중국 등 국외뿐만 아니라 국내에서도 감정 인식 분야가 중요한 과제로 연구되고 있다[1].

음성 신호는 인간의 감정 상태에 따라 억양, 음량 및 속도 등이 달라진다. 또한 전달하고자 하는 내용에 따라 특정 단어에서의 강세와 지역적인 특성이 포함된 억양 등 감정이외의 것들이 화자의 음성에 담겨져 있기 때문에, 음성에서 감정만을 따로 분리하여 분석하는데 어려움이 있다. 음성 신호를 사용하여 인간의 감정을 인식을 위해서는 감정에 따른 음성의 변화를 정확히 규명하여야 하는데, 이러한 음성과 감정의 상관관계에 대한 연구는 서구의 심리학자들과 음향학자들에 의해 먼저 이루어졌다[2,3].

* Corresponding Author : Dept. of Electrical Engineering, Kunsan National University, Korea.
E-mail: wgkim@kunsan.ac.kr
Received : March 23, 2017; Accepted : July 11, 2017

이러한 연구결과를 바탕으로 감정 인식 분야에서 다양한 연구가 진행되어 왔다[1]. Fukuda는 음성의 템포와 에너지 정보를 사용하여 여섯 가지 감정에 대한 인식 실험을 수행하였으며[4], Moriyama는 음성의 피치와 에너지 포락선을 사용하여 일본어 감정 분류 실험을 수행하였다[5]. 또한 Silva는 얼굴표정과 음성을 동시에 사용하는 바이모달 감정 인식 시스템을 사용하여 감정 인식 실험을 하였다[6]. Amol T.는 멜 켑스트럼 계수(MFCC : Mel-Frequency Cepstral Coefficients)와 네가지 음성 특성을 결합하여 감정 인식을 수행하였다[7]. 음성을 사용하여 감정 인식을 수행하기 위하여 운율, 스펙트럼, 특징 파라메터 선택 방법, 인식기 등 다양한 방법이 제안되었다[8]. Narayanan은 감정의 변화에 따라 크게 변화하는 피치로부터 감정 인식에 우수한 성능을 나타내는 파라메터를 선택하였다[9]. 또한 다양한 파라메터를 융합하여 감정 인식에 적용하는 방법도 제안되었다[10].

국내에서도 음성 및 얼굴 표정을 사용하여 감정을 인식하는 연구가 활발하게 진행되고 있다. 우리나라 전통 국악인 창에서 인간의 희로애락을 나타내는 음의 높낮이와 장단의 특성을 분석하는 연구가 수행되었고[11], 인간의 대화 내용에서 단어, 톤, 말의 빠르기나 음질 등을 분석하여 화난 감정의 특성을 파악하는 연구도 실행되었다[12]. 또한 프레임 기반의 음성 파라메터와 발화 기반의 음성 특징 파라메터를 이용한 감정 인식 실험들이 수행되었다[13]. 한국어 감정 데이터를 사용하여 피치로부터 다양한 파라메터를 생성하고 우수한 파라메터를 선택하는 연구도 진행되었다[14].

기존 연구들은 감정에 따라 변화하는 음성의 스펙트럼과 운율을 각각 사용하거나 몇 가지 파라메터를 융합하여 감정 인식에

사용하였다. 음성의 스텍트럼을 나타내는 파라메터로는 일반적으로 MFCC가 사용되었고 운율을 나타내는 파라메터로는 피치로부터 생성한 파라메터들이 사용되었다.

본 논문에서는 스펙트럼 또는 피치 파라메터만을 사용하는 기존 감정 인식 시스템의 성능을 향상하기 위하여 음성 파라메터 융합 방법을 연구하였다. 이를 위하여 기존 감정 인식 시스템에 사용된 켑스트럼 파라메터와 다양한 피치 파라메터를 융합하여 최고의 성능을 나타내는 파라메터 조합을 선정하였다. 다양한 피치 파라메터는 음성의 피치를 사용하여 수치해석적 방법과 통계적인 방법을 사용하여 생성되었다. 켑스트럼 파라메터와 결합하여 최고의 성능을 나타내는 피치 파라메터를 선정하기 위하여 가우시안 혼합 모델(GMM : Gaussian Mixture Model) 기반의 감정 인식 시스템을 대상으로 성능이 평가되었다. 파라메터 선정 방법으로는 순차적인 특징선택 방법을 사용하였다[15].

본 논문의 구성은 다음과 같다. 2장에서는 피치 파라메터에 관하여 설명하였고 3장에서는 특징 선택 방법을 사용하여 켑스트럼 계수와 피치 파라메터를 융합한 음성 감정 인식 시스템 구현 방법에 관하여 설명하였다. 4장에서는 감정 데이터베이스를 사용한 실험 과정과 결과를 다루었고 5장에서 결론을 맺었다.

2. 피치 파라메터

음성의 피치는 감정에 의해 영향을 받는 가장 중요한 요소 중의 한가지다. 이러한 피치를 감정 인식에 이용하기 위해서는 감정에 따른 피치의 변화를 파악하고 감정 인식에 사용될 수 있는 피치 파라메터를 만들어야 한다.

피치 궤적을 구하기 위하여 자기상관방법을 사용하는 Praat 소프트웨어[16]을 이용하였다. 이렇게 구한 피치 궤적를 이용하여 여러가지 파라메터를 생성하였다. 기본적으로 피치의 최대값, 최소값, 평균값, 표준편차 등을 생성하였고 통계 및 수치해석 기법을 이용하여 다양한 파라메터들을 표 1과 같이 생성하였다[9,14,17].

표 1에서 각 파라메터는 문장 또는 유성음을 기본 단위로 만들어진다. 우선 상승(rising)과 하강(falling)은 피치가 시간에 따라 각각 상승 또는 하강하는 특성을 나타내는 파라메터이다. 사분위수(interquartile)은 오름차순으로 정렬된 자료를 네 개의 구간으로 나누어, 그 중 몇 번째에 위치하는 데이터인지를 알아볼 때 사용한다. 사분범위(interquartile range)는 사분위수 3/4 위치의 값과 1/4 위치의 값의 차이를 말한다. plateaux는 피치를 1차 및 2차 미분한 값의 상관관계에 따라 정해진다. plateaux at minima는 피치를 1차 미분한 값이 0에 가깝고 2차 미분한 값이 양수일 때 그 위치의 피치 값이다. 반대로 plateaux at maxima는 피치를 1차 미분한 값이 0에 가깝고 2차 미분한 값이 음수일 때 그 위치의 피치 값이다. 피치 신호의 기울기, 곡률, 선형 또는 비선형 정도를 나타내는 파라메터로 기울기(slope), 곡률(curvature)와 변곡(inflexion)를 정의하여 1차, 2차, 3차 함수로 근사화하고 함수의 계수를 파라메터로 사용하였다. 피치 신호를 식(1)과 같이 함수로 근사화하고 가장 높은 차수의 계수 a_1, b_2, c_3를 각각 기울기, 곡률, 변곡 파라메터로 정의하였다.

$$y = a_1 x + a_0$$
$$y = b_2 x^2 + b_1 x + b_0 \quad (1)$$
$$y = c_3 x^3 + c_2 x^2 + c_1 x + c_0$$

왜도(skewness)는 분포가 평균값에 대하여 비대칭의 정도와 방향을 나타내는 값으로 비대칭도라고도 한다. 분포가 대칭이면 왜도 값은 0이고, 0보다 작으면 분포는 왼쪽으로 치우치고, 0보다 크면 분포는 오른쪽으로 치우친다. 첨도(kurtosis)는 분포의 뾰족한 정도를 나타내는 척도로, 유성음 구간의 피치 궤적의 뾰족한 정도를 나타낸다.

표 1 피치 파라메터와 단위
Table 1 Pitch parameters and its unit

parameter	unit	
	sentence	voiced
rising	O	O
falling	O	O
interquartile (range)	O	O
plateaux at maxima	O	O
plateaux at minima	O	O
slope	X	O
curvature	X	O
inflexion	X	O
skewness	X	O
kurtosis	X	O

3. 음성 감정 인식 시스템

본 연구에서는 기존 감정 인식 시스템의 성능을 향상하기 위하여 특징 선택 방법을 사용하여 피치 파라메터를 선택하고 음성 파라메터와 융합하는 방법에 관하여 연구하였다. 감정 인식에 사용된 파라메터는 멜 켑스트럼 계수와 피치로부터 구한 파라메터를 융합하는 구조를 갖는다. 이러한 감정 인식 시스템의 구조는 그림 1과 같다.

그림 1에서 음성 감정 인식 시스템은 학습과 인식의 두 단계로 이루어진다. 학습 과정에서는 감정별 학습 데이터를 사용하여 특징 파라메터를 추출한다. 여기서 사용된 특징 파라메터는 멜 켑스트럼 계수(MFCC)와 피치를 이용하여 수치 해석적 방법과 통계적인 방법을 사용하여 구한다. 다음 단계에서는 이러한 특징 파라메터를 사용하여 인식기를 학습한다. 본 연구에서는 인식기로 가우시안 혼합 모델(GMM)을 사용하였고 MFCC와 피치 파라메터에 대하여 각각 인식 모델을 생성하였다. 이렇게 생성된 모델은 인식 단계를 위해 저장된다.

인식 단계에서는 인식용 데이터베이스를 사용하여 특징 추출 단계에서 학습 단계와 동일한 방법으로 MFCC와 피치 파라메터를 구한다. 다음 단계에서는 두 가지 파라메터로 구한 각각의

GMM들을 사용하여 인식 데이터에 대한 확률 값을 각각 구한다. 마지막 결정 단계에서는 두 가지 확률 값을 융합하여 최종 확률 값을 계산한다.

MFCC와 피치 파라메터를 사용한 모델에 대한 GMM의 최종 확률 값을 융합하는 방법은 다음과 같이 수행하였다. 구별하고자 하는 S개의 감정을 감정들 $S=\{1,2,...,S\}$가 GMM들 $\lambda_1,\lambda_2,...,\lambda_N$으로 표현된다면 최종 인식 결과 $\widehat{S}$는 다음과 같이 결정된다.

$$\widehat{S}=\underset{1\le k\le S}{\arg\max}\left(\alpha p(X^{MFCC}|\lambda_k^{MFCC})+(1-\alpha)p(X^{pitch}|\lambda_k^{pitch})\right) \quad (2)$$

여기서 X^{MFCC}와 X^{pitch}는 각각 MFCC와 피치 파리메터의 벡터 열이고 λ_k^{MFCC}와 λ_k^{pitch}는 각각 학습과정에서 MFCC와 피치 파라메터를 사용하여 만든 GMM 모델이다. $p(X|\lambda)$는 입력 벡터 열 X에 대한 가우시안 혼합 확률이고 α는 가중 값으로 두 가지 확률 값의 비율을 결정한다.

순차적 특징 선택 방법(SFS)은 56개의 피치 파라메터를 순차적으로 사용하여 MFCC와 결합하였을 때 가장 우수한 성능을 나타내는 피치 파라메터 조합을 선정하였다.

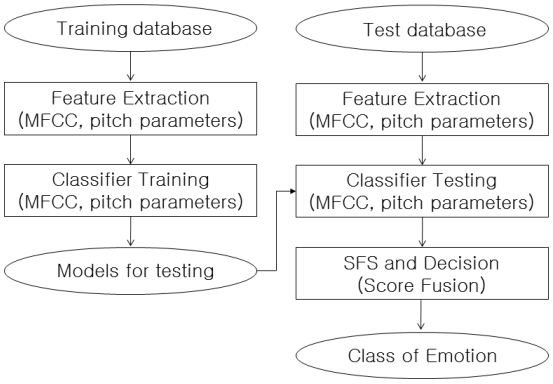

그림 1 음성 감정 인식 시스템
Fig. 1 Speech emotion recognition system

4. 실험 및 결과

4.1 데이터베이스

감정 인식 시스템의 성능 평가를 위하여 기쁨, 슬픔, 화남과 평상을 포함한 4가지 감정을 포함한 데이터베이스를 사용하였다[18]. 녹음은 평소 감정을 표현하는 훈련이 된 아마추어 연극단원 남/녀 각 15명을 대상으로 조용한 사무실 환경에서 이루어졌다. 각 화자는 45개의 서로 다른 문장을 4가지 감정으로 녹음하였다. 감정이 적절히 반영된 데이터베이스를 구축하기 위하여 주관적 평가를 수행하였다. 주관적 평가는 전체 5400개의 문장을 음성 신호처리에 숙련된 연구원 10명이 청취한 후 감정이 적절히 반영되었다고 판단되는 문장을 선택하였다. 이런 과정을 통하여 총 5400개의 음성 중에서 2237개의 음성을 선별하여 최종 데이터베이스로 구성하였다.

4.2 감정 인식 시스템의 구성

본 연구에서는 MFCC와 피치 파라메터를 융합한 감정 인식 시스템의 성능을 평가하기 위하여 GMM 기반의 화자 및 문장 독립 감정 인식 시스템을 구현하였다(그림 1).

그림 1에서 GMM 모델의 학습을 위하여 20명(남성 10명, 여성 10명)이 총 45개의 문장 중에서 35개의 문장을 녹음한 음성이 사용되었다. 인식에는 학습에 참여하지 않은 10명(남성 5명, 여성 5명)을 학습에 사용되지 않은 나머지 10개의 문장을 녹음한 음성이 사용되었다.

MFCC와 피치 파라메터를 사용하여 각각 모델을 생성하고 인식을 수행한 후에 가중 값 α를 사용하여 최종 확률을 계산한다. 이렇게 구한 확률 값을 사용하여 최종 감정인식을 수행한다.

4.3 특징 파라메터 추출

감정이 포함된 음성 신호로부터 MFCC 파라메터를 추출하여 사용하였다. MFCC 파라메터의 추출 과정은 다음과 같다. 전처리를 통하여 16KHz, 16비트로 샘플링하고, 고주파 성분을 보강한다. 이렇게 샘플링된 신호는 음성구간 검출 과정을 통해 묵음 구간을 제거한다. 검출된 음성 신호는 20ms(320샘플)의 길이를 갖는 해밍 창을 사용하여 10ms씩 이동하면서 12차의 MFCC 파라메터를 구한다. 또한 특징 파라메터의 시간적인 변화를 포함하는 ΔMFCC와 ΔΔMFCC 파라메터도 생성하였다.

감정이 포함된 음성 신호의 피치를 추정하기 위하여 Praat 소프트웨어를 이용하였다[16]. Praat 소프트웨어에서 분석 창의 크기는 40ms 로 하고 30ms씩 중첩되도록 이동하여 초당 100개의 피치 값을 계산하였다. 또한 Praat 소프트웨어의 스무딩 함수를 적용하여 피치 궤적에 포함될 수 있는 갑작스런 변화를 제거하였다.

본 실험에서 사용된 피치 파라메터를 다음과 같이 정의하였다[9,14,17]. 총 56개의 파라메터를 정의하였고 유성음과 문장 단위의 피치 파라메터를 별도로 생성하였다. 먼저 문장단위의 파라메터로 파라메터로 피치의 최대값, 최소값, 평균값, 최대 및 최소값과 평균의 차이, 최대값과 최소값 차이, 표준편차, 변곡 및 급격한 변곡의 횟수, 평균 이상과 이하에서의 평균, 상위 10%의 평균, 사분범위 등이 생성되었다. 또한 상승 및 하강 기울기의 최대값, 최소값, 평균값, 중간값, 최대 구간값, 사분위수 지속 시간 등이 생성되었다. plateaux at minima와 plateaux at maxima의 최대값, 최소값, 평균값, 중간값과 사분범위도 파라메터로 생성되었다. 왜도, 첨도, 최소 자승(least square) 파라메터도 생성되었다. 유성음 구간 단위의 파라메터는 최대값, 최소값, 변곡, 사분범위 등의 평균값들과 평균의 최대값과 최소값 파라메터 등이다. 또한 기울기, 곡률, 변곡의 평균값, 최대값, 표준편차 등도 파라메터로 생성되었다.

4.4 실험 결과

본 실험에서는 기존 감정 인식 시스템에 사용된 켑스트럼 파라메터와 다양한 피치 파라메터를 융합하여 최고의 성능을 나타내는 파라메터를 조합을 선정하였다. 우선 기존 MFCC 또는 피치 파라메터만을 사용하는 감정 인식 시스템을 대상으로 인식 실험을 실행한 후, 융합된 시스템의 성능을 평가하였다.

그림 2는 피치 파라메터와 MFCC를 각각 사용하여 감정 인식을 수행한 결과를 나타낸다. 56개의 파치 파라메터 중에서 우수한 성능을 나타내는 파라메터를 선정하기 위하여 순차적인 특징 선택 방법을 사용하였다[13]. 이러한 방법은 먼저 가장 우수한 성능의 파라메터 하나를 선정한 후, 이와 결합하였을 때 가장 우수한 성능을 나타내는 파라메터의 조합을 찾아가는 과정을 반복하여 원하는 개수의 파라메터를 선정한다. 56개의 파라메터중 15개를 조합하였을 때 가장 우수한 63.5%의 성능을 나타내었다[15]. 켑스트럼 파라메터로는 MFCC, ΔMFCC, ΔΔMFCC를 연결하여 사용하였다. 그림에서 알 수 있듯이 MFCC에 ΔMFCC와 ΔMFCC를 결합하여 사용할수록 인식 성능이 상승되어 MFCC+ΔMFCC+ΔΔMFCC의 경우에 가장 우수한 71.6%의 인식 성능을 나타내었다. 이러한 결과는 스펙트럼과 스펙트럼의 변화를 나타내는 MFCC, ΔMFCC 및 ΔΔMFCC를 사용하는 것이 피치의 통계적 특성을 나타내는 파라메터만을 사용하는 것보다 우수한 성능을 나타낸다는 것이다.

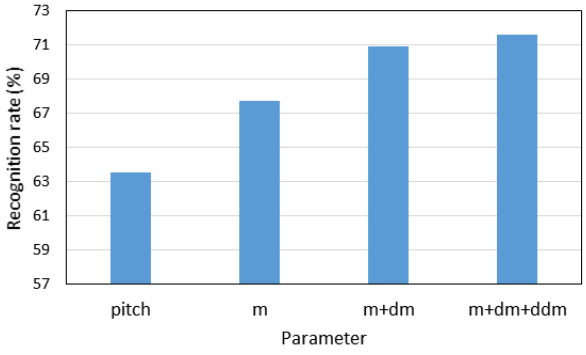

그림 2 피치와 MFCC 파라메터를 사용한 감정 인식 결과
(m : MFCC, dm : ΔMFCC, ddm : ΔΔMFCC)

Fig. 2 Emotion recognition results using pitch and MFCC
(m : MFCC, dm : ΔMFCC, ddm : ΔΔMFCC)

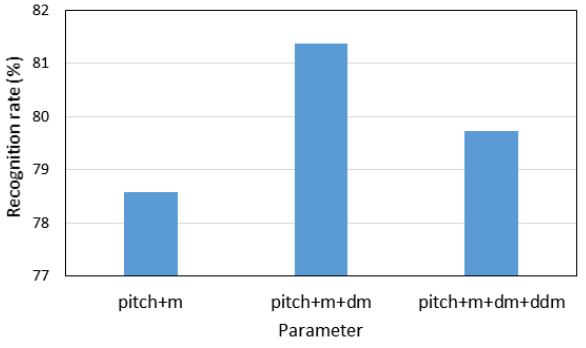

그림 3 MFCC와 피치 파라메터를 융합한 감정 인식 결과
(m : MFCC, dm : ΔMFCC, ddm : ΔΔMFCC)

Fig. 3 Emotion recognition results fusing MFCC and pitch parameter (m : MFCC, dm : ΔMFCC, ddm : ΔΔMFCC)

그림 3은 MFCC 파라메터와 피치 파라메터를 융합하고 최고의 성능을 나타내는 피치 파라메터 조합을 선정하기 위하여 GMM 기반의 감정 인식 시스템을 대상으로 순차적인 특징 선택 방법을 사용하여 인식을 수행한 결과이다. 피치 파라메터와 결합하여 가장 우수한 성능을 보인 것은 피치 파라메터와 MFCC와 ΔMFCC를 결합하여 사용하였을 때로 81.4%의 인식 성능을 나타내었다. 이는 피치만을 사용한 경우의 인식률인 63.5%보다 17.9%가 향상되었고 MFCC+ΔMFCC+ΔΔMFCC만을 사용한 경우보다 9.8%가 향상된 것으로 피치 파라메터와 MFCC 파라메터 융합의 감정 인식 성능 향상에 도움이 되는 것으로 볼 수 있다.

그림 4는 피치 파라메터와 MFCC와 ΔMFCC를 사용하고 순차적인 특징 선택 방법에 따라 선택된 피치 파라메터의 개수에 따른 인식률을 나타낸 그래프이다. 그림에서 선택된 피치 파라메터 개수가 10개 이상이면 최고값에 수렴하고 28개일 때 81.4%의 최고 인식 성능을 나타내었다. 가장 우수한 성능을 나타내는 피치 파라메터를 순서대로 나타내면 plateaux at minima의 중간값, plateaux at maxima의 사분범위, 최소 자승, 피치 상승 구간의 최대값, 피치 상승 구간의 중간값, 피치 상승 구간의 사분범위, 유성음 구간 변곡의 평균값, 피치의 표준편차, 하강 구간의 최대값, plateaux at maxima의 중간값, 평균 이하값의 평균값, 유성음 구간 곡률의 최대값, 급격한 변곡 횟수 등이다. 이렇게 선택된 파라메터들은 여러 가지 감정에 따라 피치가 급격하게 변화하는 것을 잘 표현하는 파라메터로 볼 수 있다. 이러한 결과는 피치만을 사용하는 감정 인식 시스템에서 선택된 파라메터와 유사한 것을 알 수 있다[14].

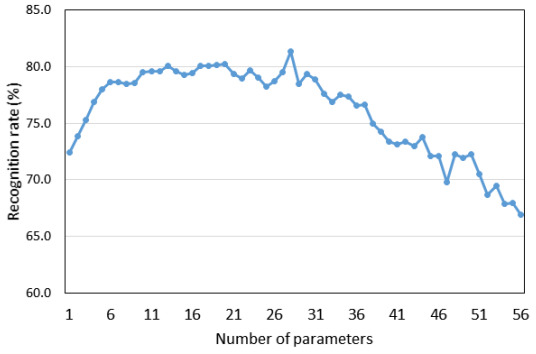

그림 4 피치 파라메터 개수에 따른 감정 인식 시스템의 성능

Fig. 4 Performance of emotion recognition system according to the number of pitch parameters

또한 식 (2)에서 피치 파라메터를 사용한 GMM의 확률 값과 MFCC를 사용한 GMM의 확률 값을 융합하기 위한 가중 값 α

는 그림 5와 같이 실험적으로 구하여졌다. 가중 값이 0.06일 때 가장 우수한 인식 성능을 나타내었다. 가중 값이 0일 때는 MFCC 만을 사용한 경우이고 1.0인 경우에는 피치만을 사용한 경우의 인식 성능을 나타낸다.

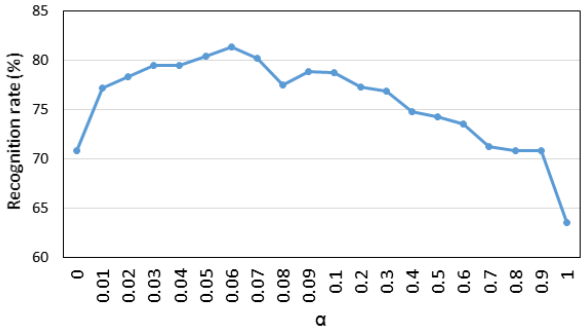

그림 5 가중 값에 따른 융합된 감정 인식 시스템의 성능

Fig. 5 Performance of fused emotion recognition system by weighted value

그림 6은 MFCC와 피치 파라메터를 단독으로 사용하는 시스템과 MFCC와 피치 파라메터를 융합하여 사용하는 시스템의 인식 성능을 비교하였다. 그림에서 MFCC+ΔMFCC+ΔΔMFCC를 사용하는 경우 71.6%, 피치 파라메터만을 사용하는 경우는 63.5%이고 피치 파라메터와 MFCC+ΔMFCC를 융합하여 사용한 경우가 81.4%로 가장 우수하였다. 이러한 값은 MFCC+ΔMFCC+ΔΔMFCC만을 사용한 경우보다 오차가 34.5% 감소하고 피치만을 사용한 경우보다 오차가 48.9% 감소한 것이다.

표 2는 피치 파라메터와 MFCC+ΔMFCC가 융합된 감정 인식 시스템의 감정별 인식률을 나타내는 표이다. 표에서 '평상' 감정의 인식은 77.6%, '기쁨' 감정의 인식은 76.0%, '슬픔' 감정의 인식은 82.4%, '화남' 감정의 인식은 89.5%로서 최종 인식률은 81.4%이다. '슬픔'과 '화남' 감정의 인식은 우수한 편이나, '평상' 감정은 '슬픔' 감정과 오인식이 많았고 '기쁨' 감정의 인식에서는 '화남' 감정과의 구분이 명확하지 않았다.

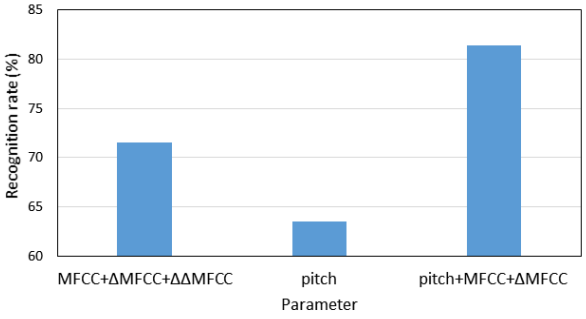

그림 6 기존 시스템과 파라메터 융합된 감정 인식 시스템의 성능 비교

Fig. 6 Performance comparison of conventional system and parameter fused emotion recognition system

표 2 파라메터 융합된 감정 인식 시스템의 인식 성능

Table 2 Recognition performance of parameter fused emotion recognition system

	recognition rate(%)			
emotion	neutral	happy	sad	angry
neutral	77.6	7.9	9.2	5.3
happy	4.0	76.0	4.0	16.0
sad	5.5	12.1	82.4	0.0
angry	3.5	7.0	0.0	89.5
average	81.4			

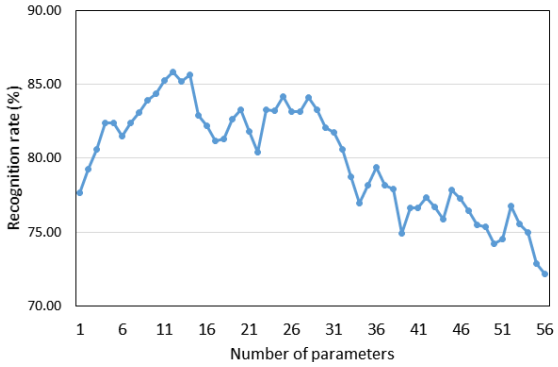

그림 7 피치 파라메터 개수에 따른 감정 검출 시스템의 성능

Fig. 7 Performance of emotion detection according to the number of pitch parameters

본 논문에서는 입력 음성을 감정이 포함되지 않은 음성과 감정이 포함된 음성으로 구분하는 감정 검출 실험을 수행하였다. 감정 검출을 위하여 기쁨, 슬픔 및 화남 데이터를 합쳐서 감정 음성(emotion)으로 재구성하였다. 데이터베이스는 평상 음성 630개와 감정 음성 1607개로 구성되었다.

그림 7은 피치 파라메터와 MFCC, ΔMFCC와 ΔΔMFCC를 사용하고 순차적인 특징 선택 방법에 따라 선택된 피치 파라메터의 개수에 따른 감성 검출 시스템의 인식률을 나타낸 그래프이다. 그림에서 선택된 피치 파라메터 개수가 12개일 때 85.9%의 최고 인식 성능을 나타내었다. 선정된 피치 파라메터는 사분범위, 급격한 변곡 횟수, 피치의 평균과 최소 차이, 평균이상 값의 평균값, 유성음 구간 변곡의 평균값, 첨도, 상승 기울기 구간의 중간값, 유성음 구간 곡률의 평균값, plateaux at maxima의 사분범위, 유성음 구간 기울기의 최대값, 하강 기울기의 사분위수 지속 구간, 유성음 구간 변곡의 표준편차 등이 선택되었다. 이러한 파라메터들은 앞에서 수행한 감정 인식에서 선택된 파라메터와 상당히 유사한 것으로, 이들 파라메터들은 여러 감정에 따른 감정 변화를 잘 나타내어 감정 검출에도 유용한 것으로 볼 수 있다.

표 3은 감정 검출 시스템의 인식 성능을 나타낸다. 표에서 감정이 없는 음성(neutral)은 인식률은 86.8%이고 감정이 포함된 음성(emotion)의 인식률은 84.8%로 최종 85.9%의 인식 성능을 나타내었다.

표 3 감정 검출 시스템의 인식 성능
Table 3 Performance of emotion detection system

	recognition rate (%)	
emotion	neutral	emotion
neutral	86.8	13.2
emotion	15.2	84.8
average	85.9	

5. 결 론

본 논문에서는 기존 감정 인식 시스템의 성능을 향상하기 위하여 켑스트럼 파라메터와 다양한 피치 파라메터를 융합하는 방법을 제안하였다. 음성의 피치를 사용하여 수치해석적 방법과 통계적인 방법을 사용하여 다양한 피치 파라메터가 생성되었다. 켑스트럼 파라메터와 결합하여 최고의 성능을 나타내는 피치 파라메터를 선정하기 위하여 GMM을 사용한 감정 인식 시스템을 대상으로 성능이 평가되었다. 파라메터 선정 방법으로는 순차적인 특징선택 방법을 사용하였다.

평상, 기쁨, 슬픔 및 화남의 4가지 감정을 구별하는 화자 및 문장 독립 감정 인식 실험에서 총 56개의 피치 파라메터중 28개를 선택하여 MFCC+ΔMFCC를 결합하여 사용한 경우가 81.4%로 가장 우수하였다. 이러한 값은 MFCC+ΔMFCC+ΔΔMFCC만을 사용한 경우보다 오차가 34.5% 감소하고 피치만을 사용한 경우보다 48.9% 감소한 것이다.

본 실험에서 특징 선택을 통하여 선택된 피치 파라메터들은 여러 가지 감정에 따라 피치가 급격하게 변화하는 것을 잘 표현하는 파라메터로 볼 수 있다. 또한 이러한 피치 파라메터들은 스펙트럼 파라메터와 융합되어 사용될 때가 각각을 독립적으로 사용될 때보다 우수한 성능을 보였다. 이러한 특성은 감정이 포함된 음성을 찾아내는 감정 검출 실험에서도 유사한 성능 향상을 나타내어 85.9%의 감정 검출성능을 나타내었다.

감사의 글

이 논문은 2012년도 정부(교육과학기술부) 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2012R1A1A4A01014421)

References

[1] R. A. Calvo, S. D'Mello, "Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications,", IEEE Trans. Affective Computing, Vol. 1, No 1, pp. 18-37, Jan 2010

[2] I. R. Murray, J. L. Arnott, "Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion", Journal Acoustical Society of America, pp.1097-1108, Feb. 1993

[3] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," IEEE Signal Process. Mag., Vol. 18, No. 1, pp. 32-80, Jan. 2001

[4] V. Kostv, S. Fukuda, "Emotion in User Interface, Voice Interaction System," IEEE International Conference on Systems, Cybernetics Representation, No.2, pp. 798-803, 2000

[5] T. Moriyama, S. Oazwa, "Emotion Recognition and Synthesis System on Speech," IEEE Intl. Conference on Multimedia Computing and System, pp. 840-844. 1999

[6] L. C. Siva, P. C. Ng, "Bimodal Emotion Recognition," Proceeding of the 4th Intl. Conference on Automatic Face and Gesture Recognition, pp. 332-335. 2000

[7] K. Amol T., R. M. R. Guddeti, "Multiclass SVM-based Language-Independent Emotion Recognition using Selective Speech Features", Proceedings of ICACCI, pp. 1069-1073, 2014

[8] R. S. Sudhkar, M. C. Anil, "Analysis of Speech Features for Emotion Detection : A review", Proceedings of 2015 International Conference on Computing Communication Control and Automation, pp. 661-664, 2015

[9] C. Busso, S. Lee, S. Narayanan, "Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection,", IEEE Trans. Speech and Audio Processing, Vol. 17, No 4, pp. 582-596, May 2009

[10] S. Ntalampiras, N. Fakotakis, "Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition", IEEE Trans. Affective Computing, Vol. 3, No. 1, pp. 116-125, Jan. 2012

[11] Y. G. Kim, Y. C. Bae, "Design of Emotion Recognition Model Using Fuzzy Logic", Proceedings of KFIS Spring Conference, 2000

[12] K. B. Sim, C. H. Park, "Analyzing the Element of Emotion Recognition from Speech", Journal of Korean Institute of Intelligent Systems, Vol. 11, No. 6, pp. 510-515, 2001

[13] N. Kim, W. Seong, H. Ha, and H. Kim, "Comparison of feature parameters for speech emotion recognition", Proceedings of Korean Institute of Communications and Information Sciences, pp. 167-168, 2016

[14] G. Lee, W. Kim, "Emotion Recognition using Pitch Parameters of Speech", Journal of Korean Institute of Intelligent Systems, Vol. 25, No. 3, pp. 272-278, June 2015

[15] P. A. Devijver, J. Kitteler, "Pattern Recognition : A

Statistical Approach", London: Prentice-Hall International, 1982
[16] P. Boersma, D. Weeninck, "PRAAT, a system for doing phonetics by computer," Inst. Phon. Sci. Univ. of Amsterdam, Amsterdam, Negherlands, Tech. Rep. 132, 1996 [Online]. Available: http://www.praat.org.
[17] D. Ververidis, C. Kotropoulos, L. Pitas, "Automatic Emotional Speech Classification", Proceedings of ICASSP'04, 2004
[18]] B. S. Kang, "Text-independent Emotion Recognition Algorithm using Speech Signal," Master thesis, Yonsei University, 2000

저 자 소 개

김 원 구 (Weon-Goo Kim)

1987년 연세대학교 전자공학과 졸업, 1989년 동 대학원 전자공학과 공학석사, 1994년 동 대학원 전자공학과 공학박사, 1998~1999년 Bell Lab, Lucent Technologies(USA) 객원연구원, 2008~2009년 Griffith University 방문교수, 1994년~현재 군산대학교 전기공학과 교수